Supporting Information

Predicting materials properties with little data using shotgun transfer learning

Hironao Yamada,[†,⊥] Chang Liu,[†,‡,⊥] Stephen Wu,[†,¶,⊥] Yukinori Koyama,[‡] Shenghong Ju,[§] Junichiro Shiomi,[‡,§] Junko Morikawa,[‡,‖] and Ryo Yoshida[*,†,‡,¶,⊥]

†*The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Tokyo 190-8562, Japan*

‡*National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan*

¶*The Graduate University for Advanced Studies, Tachikawa, Tokyo 190-8562, Japan*

§*The University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan*

‖*Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan*

⊥*Contributed equally to this work*

E-mail: yoshidar@ism.ac.jp

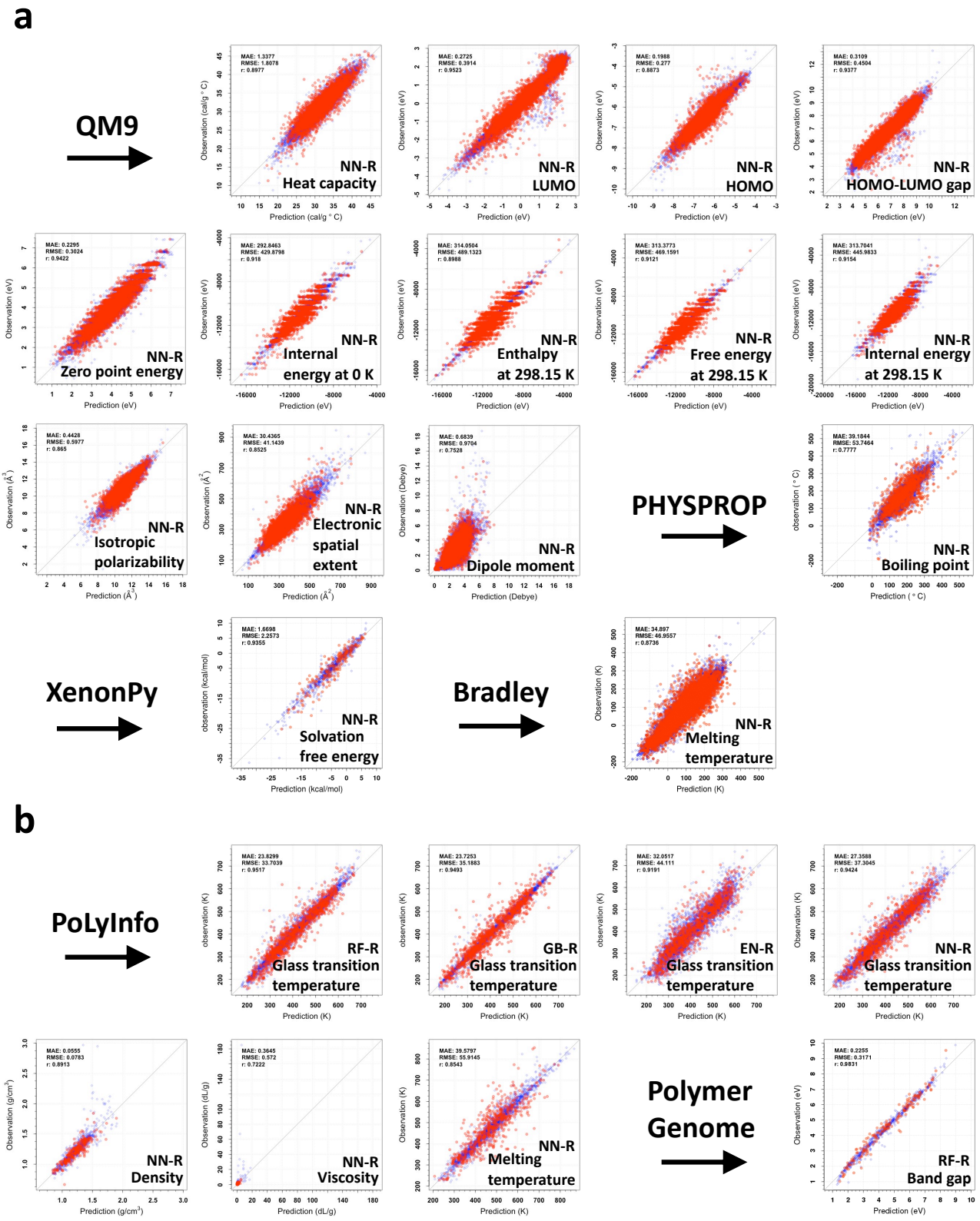
XenonPy is a Python library that implements a comprehensive set of machine learning tools for materials informatics. The current release (v0.3.7: 2019/8/7) is a prototype version, which provides some limited modules. For details, see `https://xenonpy.readthedocs.io`.

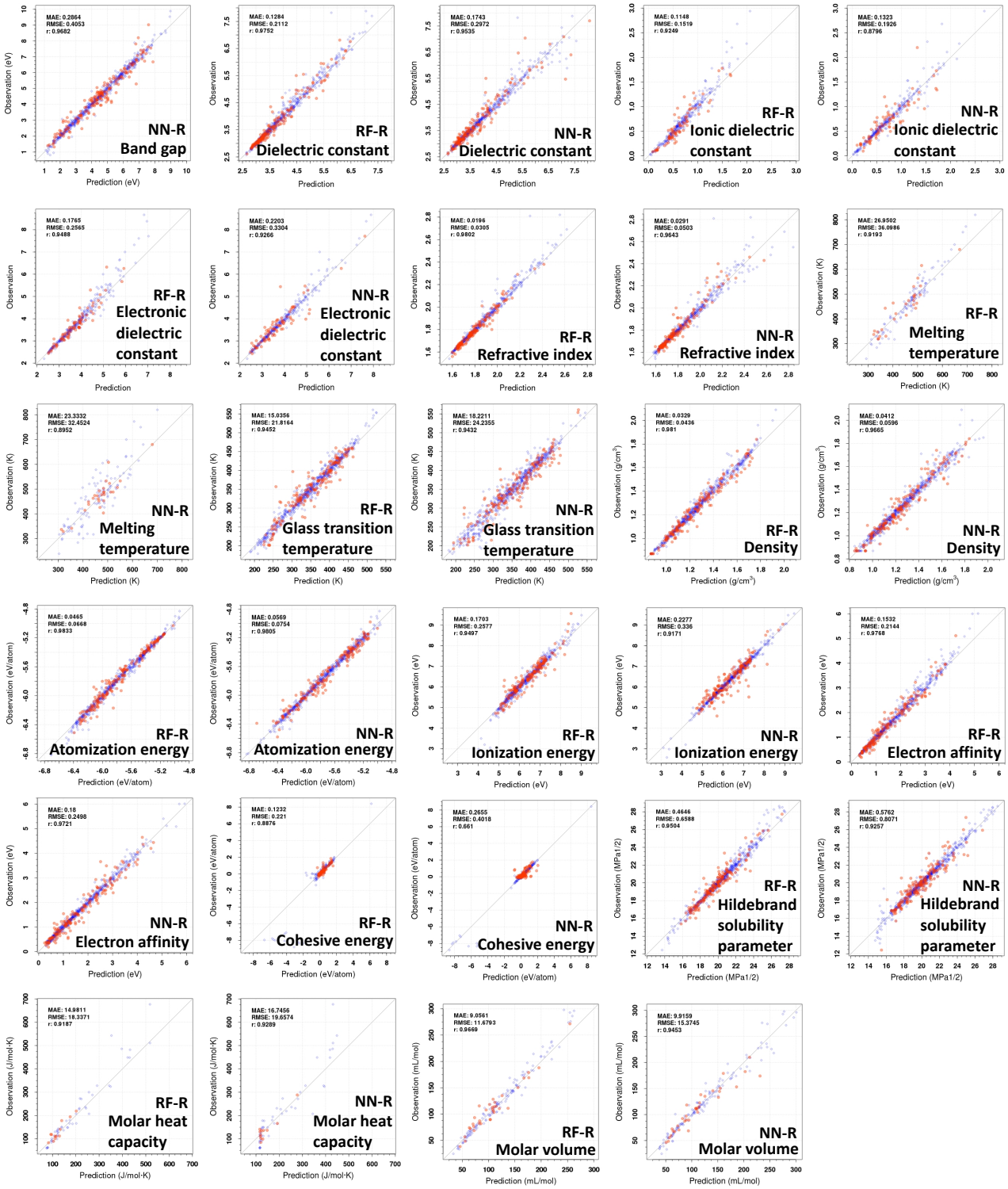
XenonPy has the following features:

- An interface with public materials databases
- A library of materials descriptors (compositional/structural/molecular descriptors)
- The XenonPy.MDL pre-trained model library (v0.1.0b, 2019/7/31: more than 140,000 models with 35 properties of small molecules, polymers, and inorganic compounds, as listed in Table 1 in the main text)
- Machine learning tools
- A transfer learning feature using pre-trained models in XenonPy.MDL

Users can interact with the search API in Python using any given query strings to obtain a specific subset of pre-trained models. Furthermore, XenonPy offers a simple-to-use tool chain for seamless performance of transfer learning using a selected pre-trained model. The full list of currently available models and sample codes (for API querying, transfer learning, and so on) is provided at `https://xenonpy.readthedocs.io/en/latest/features.html#xenonpy-mdl-and-transfer-learning`. The library is ever-growing. Examples of the prediction performance exhibited by the current best-performing models are shown in Figure S1.

Figure S1: Prediction–observation plots for current best-performing models in XenonPy.MDL. Properties of (a) small molecules, (b) polymers, and (c) inorganic compounds are ordered from left to right.




MAE: 0.2864
RMSE: 0.4053
r: 0.9682
NN-R
Band gap
Observation (eV)
Prediction (eV)
MAE: 0.1284
RMSE: 0.2112
r: 0.9752
RF-R
Dielectric constant
MAE: 0.1743
RMSE: 0.2972
r: 0.9535
NN-R
Dielectric constant
MAE: 0.1148
RMSE: 0.1519
r: 0.9249
RF-R
Ionic dielectric constant
MAE: 0.1323
RMSE: 0.1926
r: 0.8796
NN-R
Ionic dielectric constant
Observation
Prediction
MAE: 0.1765
RMSE: 0.2565
r: 0.9488
RF-R
Electronic dielectric constant
MAE: 0.2203
RMSE: 0.3304
r: 0.9266
NN-R
Electronic dielectric constant
MAE: 0.0196
RMSE: 0.0305
r: 0.9802
RF-R
Refractive index
MAE: 0.0291
RMSE: 0.0503
r: 0.9643
NN-R
Refractive index
MAE: 26.9502
RMSE: 36.0986
r: 0.9193
RF-R
Melting temperature
Observation (K)
Prediction (K)
MAE: 23.3332
RMSE: 32.4524
r: 0.8952
NN-R
Melting temperature
MAE: 15.0356
RMSE: 21.8164
r: 0.9452
RF-R
Glass transition temperature
MAE: 18.2211
RMSE: 24.2355
r: 0.9432
NN-R
Glass transition temperature
MAE: 0.0329
RMSE: 0.0436
r: 0.981
RF-R
Density
MAE: 0.0412
RMSE: 0.0596
r: 0.9665
NN-R
Density
Observation (g/cm³)
Prediction (g/cm³)
MAE: 0.0465
RMSE: 0.0668
r: 0.9833
RF-R
Atomization energy
MAE: 0.0569
RMSE: 0.0754
r: 0.9805
NN-R
Atomization energy
Observation (eV/atom)
Prediction (eV/atom)
MAE: 0.1703
RMSE: 0.2577
r: 0.9497
RF-R
Ionization energy
MAE: 0.2277
RMSE: 0.336
r: 0.9171
NN-R
Ionization energy
MAE: 0.1532
RMSE: 0.2144
r: 0.9768
RF-R
Electron affinity
Observation (eV)
Prediction (eV)
MAE: 0.18
RMSE: 0.2498
r: 0.9721
NN-R
Electron affinity
MAE: 0.1232
RMSE: 0.221
r: 0.8876
RF-R
Cohesive energy
MAE: 0.2655
RMSE: 0.4018
r: 0.661
NN-R
Cohesive energy
MAE: 0.4646
RMSE: 0.6588
r: 0.9504
RF-R
Hildebrand solubility parameter
MAE: 0.5762
RMSE: 0.8071
r: 0.9257
NN-R
Hildebrand solubility parameter
Observation (MPa1/2)
Prediction (MPa1/2)
MAE: 14.9811
RMSE: 18.3371
r: 0.9187
RF-R
Molar heat capacity
MAE: 16.7455
RMSE: 19.5574
r: 0.9289
NN-R
Molar heat capacity
Observation (J/mol·K)
Prediction (J/mol·K)
MAE: 9.0561
RMSE: 11.6793
r: 0.9669
RF-R
Molar volume
MAE: 9.9159
RMSE: 15.3745
r: 0.9453
NN-R
Molar volume
Observation (mL/mol)
Prediction (mL/mol)

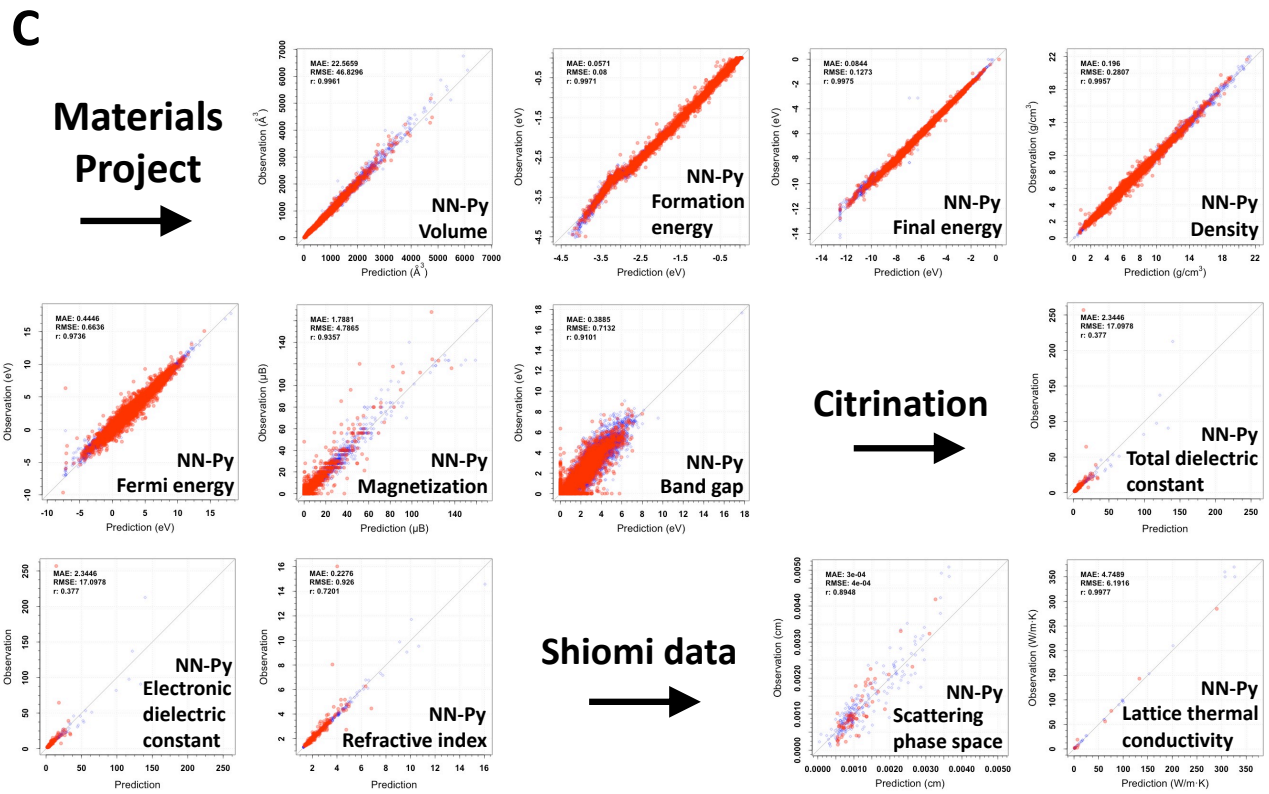


Table S1: List of fingerprint descriptors in the rcdk and RDKit libraries that were used in building the shotgun model library.

rcdk	length	RDKit	length
standard	1,024	basic fingerprints	2,048
extended	1,024	atom pairs	2,048
graph	1,024	topological torsions	2,048
hybridization	1,024	Morgan fingerprints (without feature-based)	2,048
maccs	166	Morgan fingerprints (with feature-based)	2,048
estate	79		
pubchem	881		
kr	4,860		
circular	1,024		

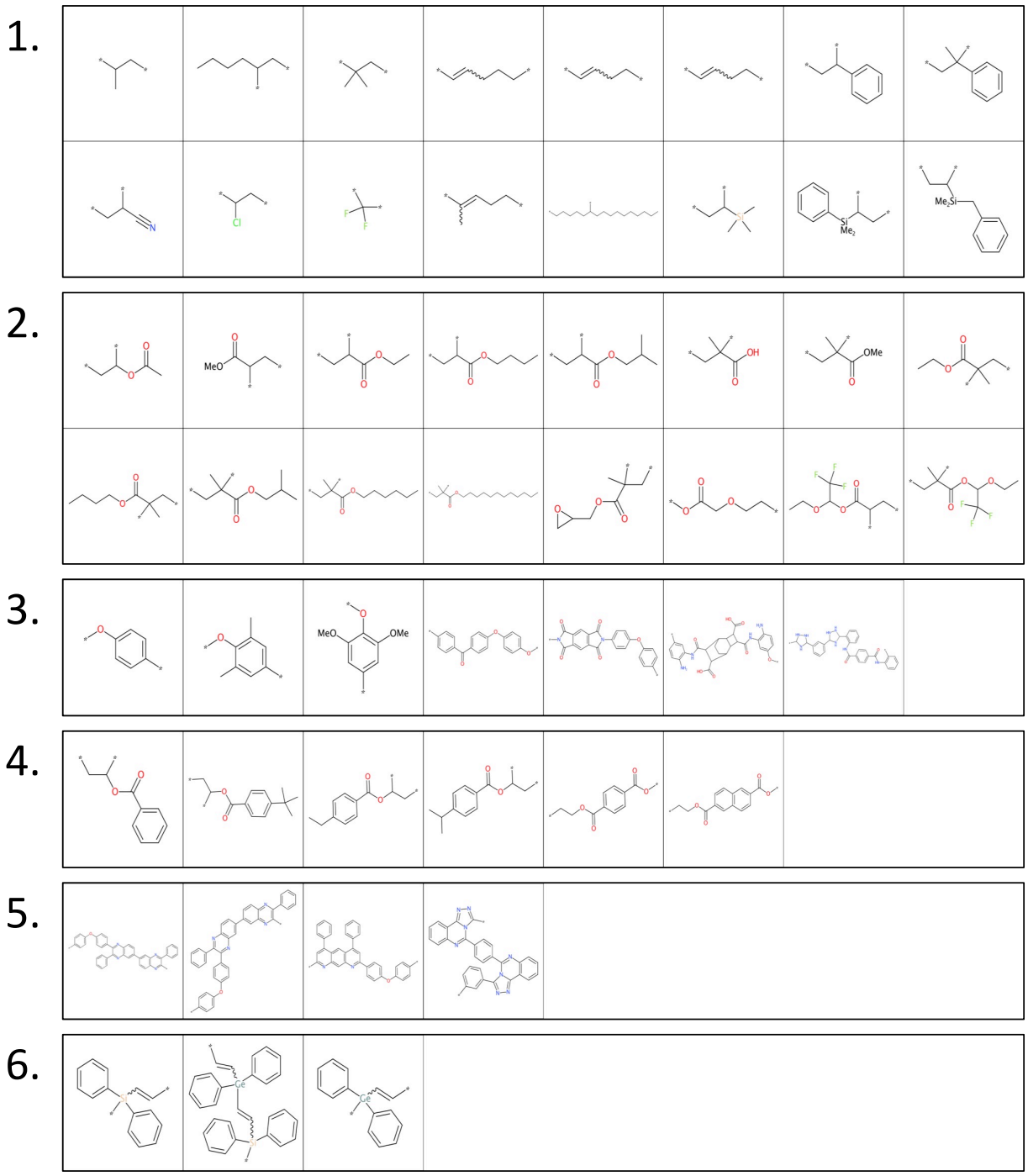


Figure S2: Chemical structures of the 52 polymers used in the task of predicting C_P. The training polymers were divided into six subgroups as numbered in the figure using the K-means clustering. Expert chemists annotated the identified clusters according to their compositional and structural features as (1) hydrocarbon mainchain polymers, (2) aliphatic esters, (3) phenols ethers, (4) aromatic esters, (5) N containing aromatics, and (6) diphenyl substituted metals. With this grouping, we performed the stratified group 6-fold CV to evaluate the generalization capability of transferred models.